Lecture 1: Optimal Prediction (with Refreshers)

36-401, Fall 2018

Thursday 6th September, 2018

Regression analysis is about investigating *quantitative*, *predictive* relationships between variables. It's about situations where there is some sort of link, tie or relation between two (or more) variables, so if we know the value of one of them, it tells us something about the other. The concrete sign of this is that knowledge of one variable lets us *predict* the other — predict the target variable better than if we didn't know the other. Pretty much everything we are going to do in this class is about crafting predictive mathematical models, seeing whether such models really have any predictive power, and comparing their predictions. Before we get into the issues of statistics and data analysis, it will help us to think what *optimal* prediction would look like, if we somehow knew all the probability distributions of all our variables.

§1 refers to many concepts from probability (reviewed in §2) and statistical inference (reviewed in §3).

1 Statistical Prediction and the Optimal Linear Predictor

1.1 Predicting a Random Variable from Its Distribution

Suppose we want to guess the value of a random variable Y. Since we don't feel comfortable with the word "guess", we call it a "prediction" instead. What's the best guess we can make?

We need some way to measure how good a guess is. Say our guess is m. The difference $Y - m$ should somehow be small. If we don't care about positive more than negative errors, it's traditional to care about the squared error, $(Y-m)^2$. Since Y is random, this will fluctuate; let's look at its expected value,

$$\mathbb{E}\left[(Y-m)^2\right] \tag{1}$$

We will call this the **mean squared error** of m, $MSE(m)$.

From the definition of variance,

$$MSE(m) = \mathbb{E}\left[(Y-m)\right]^2 = (\mathbb{E}\left[Y-m\right])^2 + \mathrm{Var}\left[Y-m\right] \tag{2}$$

The first term is the squared bias of estimating Y with m; the second term is the variance of $Y - m$. Mean squared error is bias (squared) plus variance. This is the simplest form of the **bias-variance decomposition**, which is one of the central parts of statistics.

Now remember that $\mathrm{Var}[Y - m] = \mathrm{Var}[Y]$, so

$$\begin{aligned} MSE(m) &= (\mathbb{E}[Y - m])^2 + \mathrm{Var}[Y] && (3)\\ &= (\mathbb{E}[Y] - m)^2 + \mathrm{Var}[Y] && (4) \end{aligned}$$

where the second line uses the linearity of expectations.

We would like to pick m to make this small, to minimize it (Figure 1). The variance term is irrelevant to making this small, since it's the same no matter what m is. (Remember, $\mathrm{Var}[Y]$ is about the true distribution of Y, but m is just our guess.) It should therefore play no role in the minimization.

Remember from basic calculus that one way to find the minimum[1] of a function is to take the derivative, set it to zero, and solve for the minimizing argument to the function. Here what we want to minimize is $MSE(m)$ and the argument is m, so

$$\frac{dMSE(m)}{dm} = \frac{d}{dm}\left[\mathrm{Var}[Y] + (\mathbb{E}[Y] - m)^2\right] \quad (5)$$

is the derivative we need to work out and set to zero. So, using the chain rule,

$$\frac{dMSE(m)}{dm} = \frac{d\mathrm{Var}[Y]}{dm} + 2(\mathbb{E}[Y] - m)\left(\frac{d\mathbb{E}[Y]}{dm} - \frac{dm}{dm}\right) \quad (6)$$

Changing the prediction we make, m, doesn't do anything to the true distribution of Y, so $d\mathrm{Var}[Y]/dm = d\mathbb{E}[Y]/dm = 0$, and we've got

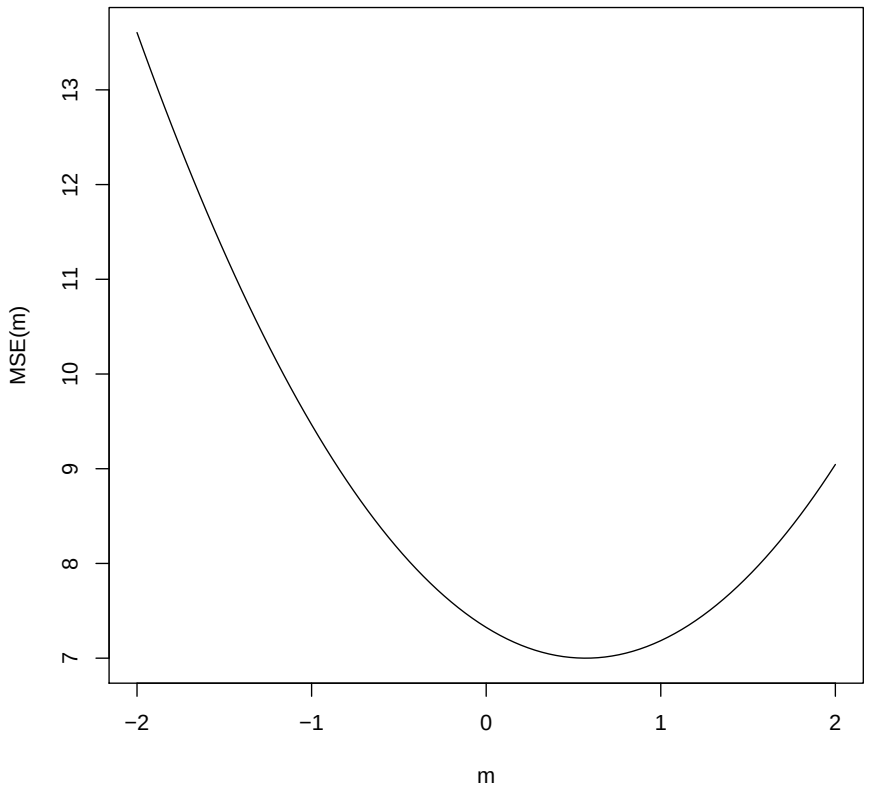
$$\frac{dMSE(m)}{dm} = -2(\mathbb{E}[Y] - m) \quad (7)$$

Say this is zero at $m = \mu$, and solve for μ:

$$\begin{aligned} -2(\mathbb{E}[Y] - \mu) &= 0 && (8)\\ \mathbb{E}[Y] - \mu &= 0 && (9)\\ \mathbb{E}[Y] &= \mu && (10) \end{aligned}$$

In other words, the best one-number guess we could make for Y is just its expected value.

[1] Or maximum; but here it's a minimum. (How could you check this, if you were worried that I was wrong?)




```
curve(7 + (0.57 - x)^2, from = -2, to = 2, xlab = "m", ylab = "MSE(m)")
```

FIGURE 1: *Mean squared error* $\mathbb{E}\left[(Y-m)^2\right]$ *as a function of the value* m *which we predict, when* $\mathbb{E}[Y] = 0.57$, $\mathrm{Var}[Y] = 7$. *(The text below the plot shows the R command used to make it.*

1.2 Predicting One Random Variable from Another

Now imagine we have two random variables, say X and Y. We know X and would like to use that knowledge to improve our guess about Y. Our guess is therefore a function of x, say $m(x)$. We would like $\mathbb{E}\left[(Y - m(X))^2\right]$ to be small.

We can use conditional expectations to reduce this problem to the one already solved.

$$\mathbb{E}\left[(Y - m(X))^2\right] = \mathbb{E}\left[\mathbb{E}\left[(Y - m(X))^2 \mid X\right]\right] \tag{11}$$

For each possible value x, the optimal value $\mu(x)$ is just the conditional mean, $\mathbb{E}[Y|X = x]$. The optimal function just gives the optimal value at each point:

$$\mu(x) = \mathbb{E}[Y \mid X = x] \tag{12}$$

This $\mu(x)$ is called the (true, optimal, or population) **regression function** (of Y on X). If we are interested in the relationship between Y and X, this is what we would really like to know, or one of the things we'd really like to know.

Unfortunately, in general $\mu(x)$ is a really complicated function, for which there exists no nice mathematical expression. The Ancestors, then, in their wisdom decided to ask "what is the best prediction we can make which is also a *simple* function of x?" In other words, they substituted a deliberately simplified *model* of the relationship for the actual relationship.

1.3 The Optimal Linear Predictor

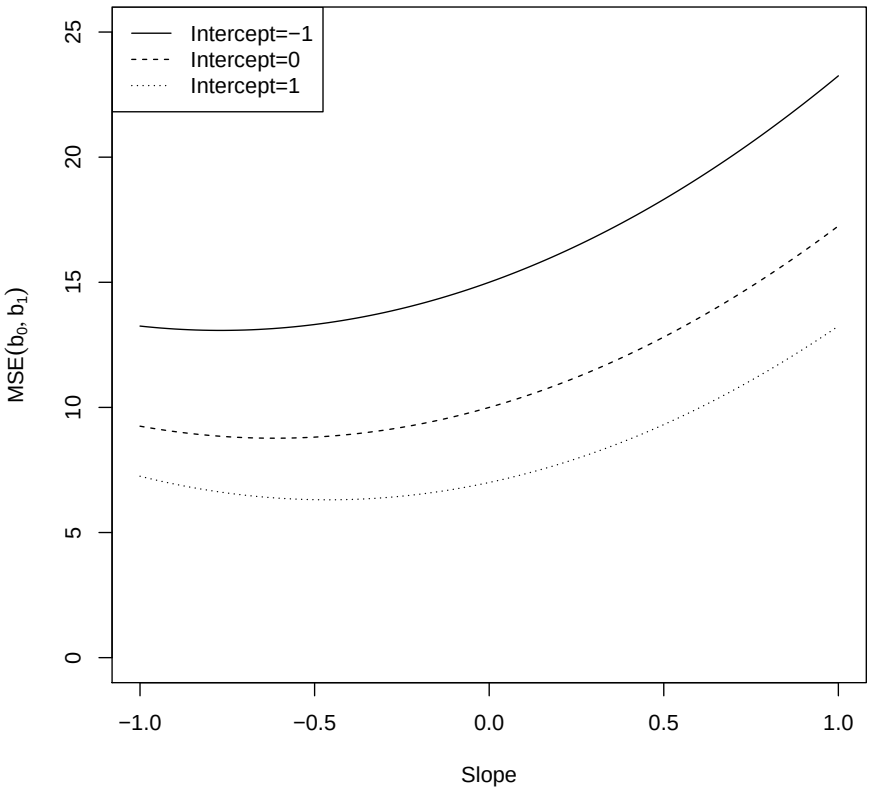
Many people regard linear functions as especially simple[2], so let us now ask "What is the optimal prediction we can make which is *linear* in X?" That is, we restrict our prediction function $m(x)$ to have the form $b_0 + b_1 x$. (To be really pedantic, that's an "affine" rather than a "linear" function.)

The mean squared error of the linear model $b_0 + b_1 x$ is now a function of two arguments, b_0 and b_1. Let's re-write it to better separate the contributions from those arguments (which we control) and the contributions from the distribution of X and Y (which are outside our control).

$$\begin{aligned}
MSE(b_0, b_1) &= \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right] && (13)\\
&= \mathbb{E}\left[Y^2\right] - 2b_0\mathbb{E}[Y] - 2b_1\mathbb{E}[XY] + \mathbb{E}\left[(b_0 + b_1 X)^2\right] && (14)\\
&= \mathbb{E}\left[Y^2\right] - 2b_0\mathbb{E}[Y] - 2b_1(\mathrm{Cov}[X, Y] + \mathbb{E}[X]\,\mathbb{E}[Y]) + b_0^2 + 2b_1\mathbb{E}[X] + b_1^2\mathbb{E}\left[X^2\right] && (15)\\
&= \mathbb{E}\left[Y^2\right] - 2b_0\mathbb{E}[Y] - 2b_1\mathrm{Cov}[X, Y] - 2b_1\mathbb{E}[X]\,\mathbb{E}[Y] + b_0^2 && (16)\\
&\quad + 2b_0 b_1\mathbb{E}[X] + b_1^2\mathrm{Var}[X] + b_1^2(\mathbb{E}[X])^2
\end{aligned}$$

(See §2 for the identities I'm using above.)

[2] Actually being precise about "how complicated is this function?" is a surprisingly hard matter. (To appreciate this, think about how a straight line may seem like a simple function, but so does a step function, and yet you need a lot of little steps to approximate a straight line...) Resolving this leads to some very deep mathematics (**??**).




```
mse <- function(b0, b1, E.Y.sq = 10, E.Y = 2, Cov.XY = -1, E.X = -0.5, Var.X = 3) {
    E.Y.sq - 2 * b0 * E.Y - 2 * b1 * Cov.XY - 2 * b1 * E.X * E.Y + b0^2 + 2 *
        b0 * b1 * E.X + Var.X * b1^2 + (E.X * b1)^2
}
curve(mse(b0 = -1, b1 = x), from = -1, to = 1, lty = "solid", ylim = c(0, 25),
    xlab = "Slope", ylab = expression(MSE(b[0], b[1])))
curve(mse(b0 = 0, b1 = x), add = TRUE, lty = "dashed")
curve(mse(b0 = 1, b1 = x), add = TRUE, lty = "dotted")
legend("topleft", legend = c("Intercept=-1", "Intercept=0", "Intercept=1"),
    lty = c("solid", "dashed", "dotted"))
```

FIGURE 2: *Mean squared error of linear models with different slopes and intercepts, when* $\mathbb{E}[X] = -0.5$, $\mathrm{Var}[X] = 3$, $\mathbb{E}[Y] = 2$, $\mathbb{E}\left[Y^2\right] = 10$, $\mathrm{Cov}[X, Y] = -1$. *Each curve represents a different intercept* b_0 *in the linear model* $b_0 + b_1 x$ *for* Y.

We minimize again by setting derivatives to zero; we now need to take two partial derivatives, which will give us two equations in two unknowns.

$$\frac{\partial \mathbb{E}\left[(Y-(b_0+b_1X))^2\right]}{\partial b_0} = -2\mathbb{E}[Y] + 2b_0 + 2b_1\mathbb{E}[X] \quad (17)$$

$$\frac{\partial \mathbb{E}\left[(Y-(b_0+b_1X))^2\right]}{\partial b_1} = -2\mathrm{Cov}[X,Y] - 2\mathbb{E}[X]\,\mathbb{E}[Y] + 2b_0\mathbb{E}[X] + 2b_1\mathrm{Var}[X] + 2b_1(\mathbb{E}[X])^2 \quad (18)$$

We'll call the optimal value of b_0 and b_1, the ones where these derivatives are exactly 0, β_0 and β_1.

The first equation is simpler, so we use it to find β_0 in terms of β_1:

$$\beta_0 = \mathbb{E}[Y] - \beta_1\mathbb{E}[X] \quad (19)$$

Some points about this equation:

- In words, it says that the optimal intercept (β_0) makes sure that the line goes through the mean Y value at the mean X value. (To see this, add $\beta_1\mathbb{E}[X]$ to both sides.)
- It's often helpful to sanity-check our math by making sure that the units balance on both sides of any equation we derive. Here, β_0 should have the same units as Y, and the right-hand side of this formula does, because β_1 has the units of Y/X.
- If the variables were "centered", with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, we'd get $\beta_0 = 0$.

Now we plug this in to the other equation:

$$0 = -\mathrm{Cov}[X,Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + \beta_0\mathbb{E}[X] + \beta_1\mathrm{Var}[X] + \beta_1(\mathbb{E}[X])^2 \quad (20)$$

$$= -\mathrm{Cov}[X,Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + (\mathbb{E}[Y] - \beta_1\mathbb{E}[X])\mathbb{E}[X] + \beta_1\mathrm{Var}[X] + \beta_1(\mathbb{E}[X])^2 \quad (21)$$

$$= -\mathrm{Cov}[X,Y] + \beta_1\mathrm{Var}[X] \quad (22)$$

$$\beta_1 = \mathrm{Cov}[X,Y]/\mathrm{Var}[X] \quad (23)$$

Some notes:

- In words, the optimal slope is the ratio between the covariance of X and Y, and the variance of X. The slope increases the more X and Y tend to fluctuate together, and gets pulled towards zero the more X fluctuates period.
- You can apply the sanity check of seeing whether this gives the right units for β_1. (Spoiler: it does.)
- The expected values $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ play no role in the formula for β_1 — only the variance and covariance matter, and they don't change when we add or subtract constants. In particular, the optimal slope doesn't change if we use instead $Y - \mathbb{E}[Y]$ and $X - \mathbb{E}[X]$.

The line $\beta_0 + \beta_1 x$ is the **optimal regression line** (of Y on X), or the **optimal linear predictor** (of Y from X).

Important Morals

1. At no time did we have to assume that the relationship between X and Y really is linear. We have derived the optimal linear approximation to the true relationship, whatever that might be.
2. The best linear approximation to the truth can be awful. (Imagine $\mathbb{E}[Y|X = x] = e^x$, or even $= \sin x$.) There is no general reason to think linear approximations *ought* to be good.
3. At not time did we have to assume anything about the marginal distributions of the variables[3], or about the joint distribution of the two variables together[4]
4. At no time did we have to assume anything about the fluctuations Y might show around the optimal regression line — that the fluctuations are Gaussian, or symmetric, or that they just add on to the regression line, etc.
5. In general, changing the distribution of X will change the optimal regression line, even if $\mathbb{P}(Y|X = x)$ doesn't change. This is because changing the distribution of X will (generally) change both $\mathrm{Cov}[X, Y]$ and $\mathrm{Var}[X]$, and the changes won't (generally) cancel out.
6. At no time did we have to assume that X came before Y in time, or that X causes Y, or that X is known precisely but Y only noisily, etc. It may be more *interesting* to model Y as a linear function of X under those circumstances, but the math doesn't care about it at all.

I will expand on that first two points a little. There is a *sort* of reason to think that linear models should work generally, which contains a kernel of truth, but needs to be used carefully.

The true regression function, as I said, is $\mu(x)$. Suppose that this is a smooth function, so smooth that we can expand it in a Taylor series. Pick then any particular value x_0. Then

$$\mu(x) = \mu(x_0) + (x - x_0)\left.\frac{d\mu}{dx}\right|_{x=x_0} + \frac{1}{2}(x - x_0)^2\left.\frac{d^2\mu}{dx^2}\right|_{x=x_0} + \ldots \quad (24)$$

Because it's tiresome to keep writing out the derivatives in this form, I'll abbreviate them as μ', μ'', etc.

[3]OK, to be pedantic, we had to assume that $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathrm{Var}[X]$ and $\mathrm{Var}[Y]$ were all well-defined and $\mathrm{Var}[X] > 0$.

[4]Except, to keep being pedantic, that $\mathrm{Cov}[X, Y]$ was well-defined.

For x close enough to x_0, we can get away with truncating the series at first order,

$$\mu(x) \approx \mu(x_0) + (x - x_0)\mu' \tag{25}$$

and so we could identify that first derivative with the optimal slope β_1. (The optimal intercept β_0 would depend on $\mu(x_0)$ and the distribution of $x - x_0$.) How close is enough? Close enough that all the other terms don't matter, so, e.g., the quadratic term has to be negligible, meaning

$$|x - x_0|\mu' \gg |x - x_0|^2\mu''/2 \tag{26}$$

$$2\mu'/\mu'' \gg |x - x_0| \tag{27}$$

Unless the function is really straight, therefore, any linear approximation is only going to be good over very short ranges.

It *is* possible to do a lot of "local" linear approximations, and estimate $\mu(x)$ successfully that way — in fact, we'll see how to do that in 402 (or read **?**). But a justification for a *global* linear model, this is weak.

A better justification for using linear models is simply that they are *computationally* convenient, and there are many situations where computation is at a premium. If you have huge amounts of data, or you need predictions very quickly, or your computing hardware is very weak, getting a simple answer can be better than getting the *right* answer. In particular, this is a rationale for using linear models to make *predictions*, rather than than for caring about their *parameters*.

All of that said, we are going to spend most of this course talking about doing inference on the parameters of linear models. There are a few reasons this is not *totally* perverse.

- The theory of linear models is a special case of the more general theory which covers more flexible and realistic models. But precisely because it is such a special case, it allows for many simplifying short-cuts, which can make it easier to learn, especially without advanced math. (We can talk about points and lines, and not about reproducing-kernel Hilbert spaces.) Learning linear models first is like learning to swim in a shallow pool, rather than in the ocean with a gorgeous reef, deceptive currents, and the occasional shark. (By the end of the year, you will know how to dive with small sharks.)

- Because linear models are so simple, for most of the last two hundred odd years they were the only sort of statistical model people could actually *use*. This means that lots of applications of statistics, in science, in policy and in industry, has been done on linear models. It also means that lots of consumers of *statisticians*, in science, in policy and in industry, expect linear models. It is therefore important that you understand thoroughly both how they work and what their limitations are.

Throughout the rest of the course, we are going to tack back and forth between treating the linear model as exactly correct, and treating it as just a

FIGURE 3: *Statistician (right) receiving population moments from the Oracle (left).*

more-or-less convenient, more-or-less accurate approximation. When we make the stronger assumption that the linear model is right, we will be able to draw stronger conclusions; but these will not be much more secure than that assumption was to start with.

1.4 Probability versus Statistics

Everything I've gone over so far is purely mathematical. We have been pretending (Figure 3) that we have gone to the Oracle, and in a mystic trance they have revealed to us the full joint probability distribution of X and Y, or at least the exact values of all of their moments. As in many mathematical problems, therefore, we have idealized away everything inconvenient. In reality, we never know the full probability distribution of the variables we are dealing with[5]. Rather than exact knowledge from the Oracle, we have only a limited number of noisy samples from the distribution. The *statistical* problem is that of drawing inferences about the ideal predictor from this unpromising material.

2 Reminders from Basic Probability

The expectation (or expected value) of a continuous random variable X with probability density function $p(x)$ is

$$\mathbb{E}[X] = \int xp(x)dx \tag{28}$$

while the expectation of a discrete random variable with probability mass function $p(x)$ is

$$\mathbb{E}[X] = \sum_x xp(x) \tag{29}$$

[5]Even the idea that the variables we see *are* randomly generated from a probability distribution is a usually-untestable assumption.

(Because everything is parallel for the discrete and continuous cases, I will not keep writing out both forms; after tossing a coin, I will just write out the integrals.)

The expectation of any function of a random variable $f(X)$ is

$$\mathbb{E}\left[f(X)\right] = \int f(x)p(x)dx \tag{30}$$

(Of course, $f(X)$ has its own distribution, with a density we might call p_f; can you prove that that $\int f(x)p(x)dx = \int hp_f(h)dz$?)

$X - \mathbb{E}\left[X\right]$ is the **deviation** or **fluctuation** of X from its expected value.

The **variance** of X is

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] \tag{31}$$

The **covariance** of X and Y is

$$\mathrm{Cov}\left[X, Y\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] \tag{32}$$

The covariance is positive when X and Y tend to be above or below their expected values together, and negative if one of them having a positive fluctuation tends to go with the other having a negative fluctuation.

2.1 Algebra with Expectations, Variances and Covariances

We're going to deal a lot with expectation values, variances and covariances. There are some useful bits of algebra about these, which I will now remind you of. You will commit them to memory (either deliberately or because you'll use them so often).

1. *Linearity of expectations*

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right] \tag{33}$$

2. *Variance identity*

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] \tag{34}$$

3. *Covariance identity*

$$\mathrm{Cov}\left[X, Y\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] \tag{35}$$

4. *Covariance is symmetric*

$$\mathrm{Cov}\left[X, Y\right] = \mathrm{Cov}\left[Y, X\right] \tag{36}$$

5. *Variance is covariance with itself*

$$\mathrm{Cov}\left[X, X\right] = \mathrm{Var}\left[X\right] \tag{37}$$

6. *Variance is not linear*

$$\mathrm{Var}\left[aX+b\right] = a^2\mathrm{Var}\left[X\right] \tag{38}$$

7. *Covariance is not linear*

$$\mathrm{Cov}\left[aX+b, Y\right] = a\mathrm{Cov}\left[X, Y\right] \tag{39}$$

8. *Variance of a sum*

$$\mathrm{Var}\left[X+Y\right] = \mathrm{Var}\left[X\right] + \mathrm{Var}\left[Y\right] + 2\mathrm{Cov}\left[X, Y\right] \tag{40}$$

9. *Variance of a big sum*

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left[X_i, X_j\right] = \sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] + 2\sum_{i=1}^{n-1}\sum_{j>i}\mathrm{Cov}\left[X_i, X_j\right] \tag{41}$$

10. *Law of total expectation*

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right] \tag{42}$$

    Remember: $\mathbb{E}\left[Y|X\right]$ is a function of X; it's random.

11. *Law of total variance*

$$\mathrm{Var}\left[Y\right] = \mathrm{Var}\left[\mathbb{E}\left[Y|X\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[Y|X\right]\right] \tag{43}$$

12. *Independence implies zero covariance* If X and Y are independent, $\mathrm{Cov}\left[X, Y\right] = 0$. The reverse is *not* true; $\mathrm{Cov}\left[X, Y\right] = 0$ is even compatible with Y being a function of X.

2.2 Convergence

The Law of Large Numbers Suppose that $X_1, X_2, \ldots X_n$ all have the same expected value $\mathbb{E}\left[X\right]$, the same variance $\mathrm{Var}\left[X\right]$, zero covariance with each other. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}\left[X\right] \tag{44}$$

In particular, if the X_i all have the same distribution and are independent ("independent and identically distributed", IID) then this holds.

Note: There are forms of the law of large numbers which don't even require a finite variance, but they are harder to state. There are also ones which do not require constant means, ore even a lack of covariance among the X_i, but they are also harder to state.

Central limit theorem If the X_i are IID, then as $n \to \infty$, the distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$ approaches $\mathcal{N}(\mathbb{E}[X], \mathrm{Var}[X]/n)$, regardless of the distribution of the X_i.

Mathematically, it is nicer to have the limit that we're converging to not change with n, so this is often stated as

$$\sqrt{n}\frac{\overline{X}_n - \mathbb{E}[X]}{\mathrm{Var}[X]} \rightsquigarrow \mathcal{N}(0,1) \tag{45}$$

Note: There are versions of the central limit theorem which do *not* assume independent or identically distributed variables being averaged, but they are considerably more complicated to state.

3 Reminders from Basic Statistics: Estimation

We observe values $X_1, X_2, \ldots X_n$ from some distribution. We don't know the distribution, so we imagine writing it down with one or more unknown parameters, $f(x;\theta)$. A **statistic** is a function of the data, and the data alone. An **estimator** is a statistic which takes a guess at the parameter θ, or some function of it, $h(\theta)$. (For instance we might want to estimate $\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$.) We will generically write such an estimator as $\hat{\theta}_n$, with the hat to distinguish it from the true value of the parameter, and the subscript n to emphasize that it will change as we get more data.

An estimator is a random variable; it inherits its distribution from that of the data X_i. This is often called the **sampling distribution** of the estimator.

An estimator is **consistent** if $\hat{\theta}_n \to \theta$, whatever the true θ might be. An estimator which is not consistent is **inconsistent**, and usually not very good.

The **bias** of an estimator is $\mathbb{E}\left[\hat{\theta}_n - \theta\right] = \mathbb{E}\left[\hat{\theta}_n\right] - \theta$. An estimator is **unbiased** if its bias is zero for all θ.

An estimator also has a variance, $\mathrm{Var}\left[\hat{\theta}\right]$. The **standard error** of an estimator is its standard deviation, the square root of the variance. We give it the name "standard error" to remind ourselves that this is telling us about how precise our estimate is. N.B., there are more standard errors than just the standard error in the mean (see below).

An estimator cannot be consistent unless its standard error goes to zero as n grows. If both the standard error and the bias go to zero, that guarantees consistency, but there are exceptional circumstances where asymptotically biased estimators are still consistent.

Example: Sample Mean The expectation value $\mathbb{E}[X]$ is either a parameter of a distribution, or a function of the parameters. The sample mean $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ is a statistic, since it is a function of the data alone. The sample mean can be used as an estimator of $\mathbb{E}[X]$, and is a natural choice for this role. If the X_i are IID, then the law of large numbers tells us that $\overline{X}_n$ is

a consistent estimator of $\mathbb{E}[X]$. The central limit theorem tells us that the sampling distribution is asymptotically Gaussian.

It is easy to prove (so do so!) that $\mathbb{E}\left[\overline{X}_n\right] = \mathbb{E}[X]$, hence the mean is an unbiased estimator of the expected value. Notice that $\mathrm{Var}\left[\overline{X}_n\right] = \mathrm{Var}[X_1]/n$, which, as promised above, goes to zero as $n \to \infty$. The corresponding standard deviation is $\sigma/\sqrt{n}$, which is the "standard error in the mean". (Again, every estimator of every quantity has its own standard error, which is not just this.)

Example: Shrunken Sample Mean As an alternative estimator, consider $\frac{n}{n+\lambda}\overline{X}_n$, where you get to set the number $\lambda > 0$ (but then you have to use the same λ for all n). You should be able to convince yourself that (i) at every n and every λ, it has a strictly smaller variance than $\overline{X}_n$, and hence a strictly smaller standard error; (ii) it is a biased estimator of $\mathbb{E}[X]$, with a bias which depends on $\mathbb{E}[X]$, λ and n; (iii) for every $\mathbb{E}[X]$ and λ, the bias goes to zero as $n \to \infty$; (iv) it is a consistent estimator of $\mathbb{E}[X]$. This is an example of what is called a "shrinkage" estimator, where the obvious estimate is "shrunk" towards zero, so as to reduce variance.